A Study of Irrelevant Variability Normalization Based Training and Unsupervised Online Adaptation for LVCSR

Guangchuan Shi[1,2], *Yu Shi*[1], *Qiang Huo*[1]

[1]Microsoft Research Asia, Beijing, China
[2]MOE-Microsoft Key Laboratory for Intelligent Computing and Intelligent Systems
[2]Department of Computer Science and Engineering, Shanghai Jiao Tong University, Shanghai, China
sgc1984@sjtu.edu.cn, yushi@microsoft.com, qianghuo@microsoft.com

NTNU

Introduction

- Irrelevant variability normalization (IVN) has been proposed for acoustic modeling, training and adaptation

Training

Testing

Approach

- Feature Transformation Function

$$x_t = F(y_t;\Theta) = A^{(e_t)} y_t + b^{(l_t)}$$

  - y_t is the *t*-th original *D*-dimensional feature vector
  - x_t is the transformed feature vector
  - $A^{(e_t)}$ is the $D \times D$ nonsingular transformation matrix
  - $b^{(l_t)}$ is the *D*-dimensional bias vector
  - e_t and l_t are the labels informed by "Acoustic Sniffing" module
  - $\Theta = \{A^{(e)}, b^{(l)} \mid e = 1,2,\cdots,E; l = 1,2,\cdots,L\}$

Approach (Cont.)

- Moving-Window Approach to Acoustic Sniffing
  - In training stage, given the feature vector sequences of training data, for the *t*-th frame of raw feature vector y_t , we first calculate 6 new D-dimensional feature vectors

$$
\begin{aligned}
\overline{y}_{t-3} &= \frac{1}{4}(y_{t-9} + y_{t-8} + y_{t-7} + y_{t-6}) \\
\overline{y}_{t-2} &= \frac{1}{3}(y_{t-5} + y_{t-4} + y_{t-3}) \\
\overline{y}_{t-1} &= \frac{1}{2}(y_{t-2} + y_{t-1}) \\
\overline{y}_{t+1} &= \frac{1}{2}(y_{t+1} + y_{t+2}) \\
\overline{y}_{t+2} &= \frac{1}{3}(y_{t+3} + y_{t+4} + y_{t+5}) \\
\overline{y}_{t+3} &= \frac{1}{4}(y_{t+6} + y_{t+7} + y_{t+8} + y_{t+9})
\end{aligned}
$$

$\Rightarrow$ concatenate into a single super vector z_t

  - Given the new set of training feature vectors $\{z_t\}$, a Gaussian mixture model (GMM) with *K* components is trained

Approach (Cont.)

- Then, two hierarchical trees can be constructed by using a divisive Gaussian clustering method with *E* and *L* leaf nodes
  - to form two Gaussian codebooks

$$\left\{N\left(z;\xi_e^{(A)},R_e^{(A)}\right)\middle|\, e=1,2,\cdots,E\right\}$$

$$\left\{N\left(z;\xi_l^{(b)},R_l^{(b)}\right)\middle|\, l=1,2,\cdots,L\right\}$$

- In both IVN-based training and recognition stage, a label can be assigned for transformation matrix and bias vector

$$e_t=\arg\max_e N\left(z;\xi_e^{(A)},R_e^{(A)}\right)$$

$$l_t=\arg\max_l N\left(z;\xi_l^{(b)},R_l^{(b)}\right)$$

Approach (Cont.)

- IVN-based ML Training
  - Assume each basic speech unit is modeled by a Gaussian mixture continuous density hidden Markov models (CDHMM) whose parameters are denoted as

$$\lambda = \{\pi_s, a_{ss'}, c_{sm}, \mu_{sm}, \Sigma_{sm}\}$$

  - Let $\Lambda = \{\lambda\}$ denote the set of CDHMM parameters and $\mathcal{Y} = \{Y_i | i = 1, 2, \cdots I\}$ the set of training data

  - By using the acoustic sniffing technique, two sets of frame labels for transformation matrix and bias vector $\mathcal{E}$ and $\mathcal{L}$ derived from $\mathcal{Y}$.

Approach (Cont.)

- The IVN-based ML training is to maximize, by adjusting feature transform parameters Θ and HMM parameters Λ ,the following likelihood function

$$F(\Theta, \Lambda) = \prod_{i=1}^{I} p(Y_i|\Theta, \Lambda, \mathcal{E}, \mathcal{L})$$

- They used *method of alternating variables* to maximize the above objective function

- **Step 1:** *Initialization*
  - The set of HMM parameters is initialized as the one trained using a traditional ML training approach
  - The feature transformation matrices are initialized as identity matrices and the bias vectors are initialized as zero vectors

Approach (Cont.)

- **Step 2:** *Estimate Feature Transformation Parameters Θ by Fixing HMM Parameters Λ*
  - Given the fixed HMM parameters $\overline{\Lambda}$, the likelihood function $F\left(\Theta, \overline{\Lambda}\right)$ can be increased by running several EM iterations to re-estimate Θ

$$\mathcal{Q}(\Theta, \overline{\Theta}) = \sum_{t.s.m} \gamma_{sm}(t) \log p_{sm}(y_t|\Theta, \overline{\Lambda})$$

$$p_{sm}(y_t|\Theta, \overline{\Lambda}) = \mathcal{N}(\mathcal{F}(y_t; \Theta); \overline{\mu}_{sm}, \overline{\Sigma}_{sm}) |\det(A^{(e_t)})|$$

- **Step 3:** *Estimate HMM Parameters Λ by Fixing Feature Transformation Parameters Θ*

- **Step 4**: *Repeat Step 2 and Step 3 N_C times*

Approach (Cont.)

- Unsupervised Online Adaptation
  - **Step 1:** Transform $F(y_t;\Theta)$ with pretrained transform parameters. Do first-pass recognition by using generic HMMs
  - **Step 2:** Given the recognized transcription the transform parameters are re-estimated.
  - **Step 3:** Transform *Y* with the *updated* parameters $\hat{\Theta}$. Do recognition by using generic HMMs
  - **Step 4:** Repeat Step 2 and Step 3 until a pre-specified criterion is satisfied

NTNU

Experiments and Results

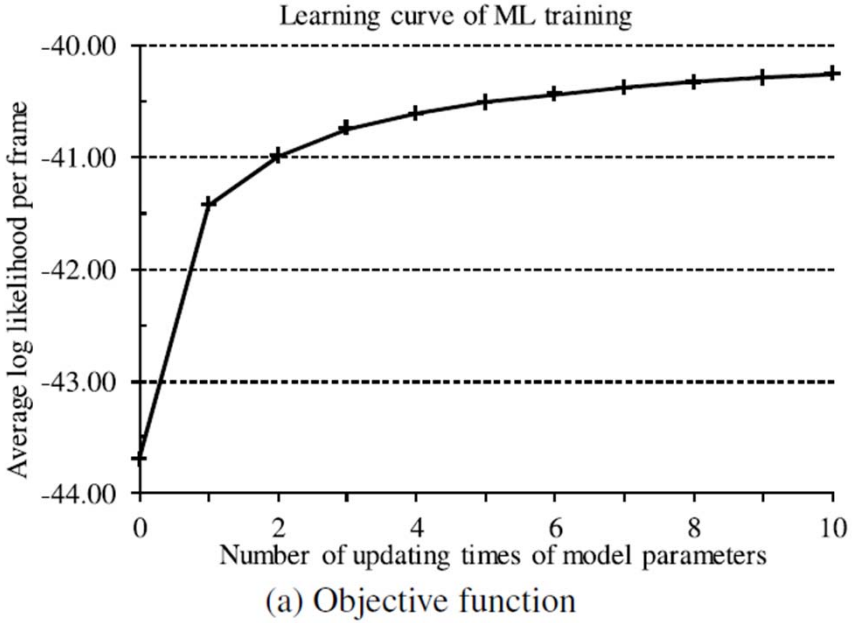
- Experimental Setup
  - The speech corpus is Switchboard-1 corpus
    - 4870 sides of conversations (about 300 hours speech) from 520 speakers as training data
    - 40 sides of Switchboard-1 conversations (about 2 hours speech) from the 2000 Hub5 evaluation as testing data
  - For feature extraction in front-end, they used 39 PLP_E_D_A. Conversation-side based mean and variance normalization was applied in both training and recognition stages
  - For acoustic modeling, they used phonetic decision-tree based tied-state triphone CDHMMs with 9302 states and 40 Gaussian components per state
  - The recognition vocabulary contains 22641 unique words

NTNU

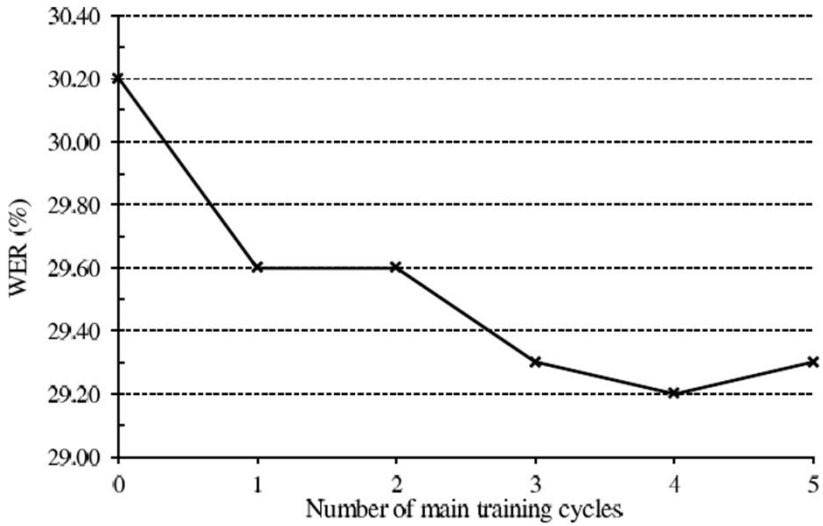
Experiments and Results

- In moving-window based acoustic sniffing, the setting of relevant control parameters is as follows: *K = 1024*, *E* = 8, *L = 8.*

- Learning Behavior of IVN-based ML Training



(a) Objective function



(b) Word error rate on testing set

- The WER is reduced from 30.2% for baseline system to 29.3% after 5 main cyles of IVN-based training

Experiments and Results

- Effects of Unsupervised Online Adaptation

Table 1: *Comparison of different approaches.*

Method	WER (%)	Relative (%)
Baseline	30.2	N/A
Baseline + MLLR	28.4	5.96
CMLLR-based AT	29.5	2.32
CMLLR-based AT + OLA	27.5	8.94
IVN-based Training	29.3	2.98
IVN + OLA	27.2	9.93

  - After two cycles of recognition and OLA, the WER is reduced from 29.3% to 27.2%
  - Apparently IVN-based approach achieves the best performance

Conclusion and Discussions

- The IVN-based approach has at least the following advantages
  - The open mechanism of acoustic sniffing offers new opportunities and flexibility for innovation
  - Because IVN-based approach can be implemented as a feature transformation approach, no change of speech decoder has to be made
- Ongoing and future works for IVN-based framework include
  - explore different acoustic sniffing techniques
  - investigate the effectiveness of using discriminative training for both transforms and generic HMM parameters for LVCSR
  - investigate the effectiveness of a hybrid approach for LVCSR

NTNU